Wiggle150: Benchmarking Density Functionals And Neural Network Potentials On Highly Strained Conformers

Rebecca R. Brew[1,‡], Ian A. Nelson[1, ‡], Meruyert Binayeva1, Amlan S. Nayak[1], Wyatt J. Simmons[1], Joseph J. Gair[1,], Corin C. Wagen[2,*]*

[1] Michigan State University, Lansing MI, 48824

[2] Rowan Scientific, Boston MA 02134

ABSTRACT Accurate benchmarks are key to assessing the accuracy and robustness of computational methods, yet most available benchmark sets focus on equilibrium geometries, limiting their utility for applications involving non-equilibrium structures such as ab initio molecular dynamics and automated reaction-path exploration. To address this gap, we introduce Wiggle150, a benchmark comprising 150 highly strained conformations of adenosine, benzylpenicillin, and efavirenz. These geometries—generated via metadynamics and scored using DLPNO-CCSD(T)/CBS reference energies—exhibit substantially larger deviations in bond lengths, angles, dihedrals, and relative energies than other conformer benchmarks. We evaluate a diverse array of computational methods, including density-functional theory, composite quantum chemical methods, semiempirical models, neural network potentials, and force fields, on predicting relative energies for this challenging benchmark set. The results highlight multiple

methods along the speed–accuracy Pareto frontier and identify AIMNet2 as particularly robust among the NNPs surveyed. We anticipate that Wiggle150 will be used to validate computational protocols involving non-equilibrium systems and guide the development of new density functionals and neural network potentials.

Introduction

The development of standardized, high-quality benchmarks is crucial to the continual advance of scientific methods. In atomistic simulation, benchmarking drives the development of new density functionals, basis sets, & corrections and allows researchers to assess the accuracy of new machine-learning-based approaches. Following the oft-cited adage that "you get what you screen for," [1] ensuring that good benchmarks exist for all desired applications of atomistic simulation is key to driving rigorous and systematic progress in computational chemistry.

Most quantum-chemical benchmarks focus exclusively on equilibrium structures, i.e. structures for which the net force on every atom is zero. For instance, the popular GMTKN55[2] collection of main-group thermochemistry benchmarks is entirely composed of ground- and transition-state structures, and even Korth and Grimme's exotic "mindless" benchmark sets only include completely optimized structures.[3] Exceptions are generally limited to structures with a single non-equilibrium degree of freedom, like the S66x8[4,5] benchmark set of non-covalent dimer dissociation curves. Since many emerging applications of quantum chemistry involve structures where multiple internal coordinates are far from equilibrium, like ab initio molecular dynamics & metadynamics[6,7] and automated reaction-path exploration, the lack of non-equilibrium benchmarks makes it challenging to assess the effect of different theoretical methods on these important workflows.

Good benchmarks for non-equilibrium structures are also important for assessing the robustness of neural network potentials (NNPs). Today, the majority of NNPs are trained on equilibrium or near-equilibrium structures,[8] which leads them to perform poorly on high-energy structures that are underrepresented or omitted in their training data.[9,10] Many pretrained models also display catastrophic failure to generalize to unseen configurations, causing instability in molecular-dynamics simulations even when overall force-based errors are low.[11] Unfortunately, detecting this instability currently requires running long molecular-dynamics simulations and watching for unexpected behavior, which is time-consuming and difficult to reproduce. We reasoned that intentionally creating a benchmark set of high-energy conformations might enable informed comparisons of the robustness of different NNP models and architectures without the stochasticity inherent to molecular dynamics-based benchmarks.

Here, we report the Wiggle150 benchmark set, which comprises geometries and DLPNO-CCSD(T)/CBS energies for three different molecules with 50 strained conformations each. When compared to existing conformational benchmarks, Wiggle150 displays markedly higher variation in bond lengths, angles, dihedrals, and relative energies. We use Wiggle150 to study the robustness of different computational methods, including density-functional theory, "composite" quantum chemical methods, semiempirical methods, and neural network potentials, and make recommendations for methods at many different points along the speed–accuracy Pareto frontier.

Methods

Benchmark Set. We selected three organic molecules for detailed investigation in this study: adenosine, benzylpenicillin, and efavirenz. These structures were chosen because of their biological relevance and because they represented a relatively wide variety of functional groups:

arenes, amides, alcohols, carboxylic acids, thioethers, aryl halides, trifluoromethyl groups, alkynes, basic N-heterocycles, and 3-, 4-, 5- & 6-membered rings.

We generated initial structures for each molecule using Rowan[12] and ran 10 ps of metadynamics[13] using GFN2-xTB[14] with a 1 fs timestep. (The mass of the hydrogen atoms was kept as 1 amu, and the default SHAKE constraints were disabled.) From each output trajectory, 50 dissimilar conformations were selected by agglomerative clustering on heavy-atom RMSD. The ground-state conformer was identified by running a conformational search in Rowan ("rapid" mode, which employs the ETKDG[15,16] algorithm) and optimizing the lowest-energy conformer at the B3LYP-D3BJ/def2-TZVP level of theory.

To ensure the suitability of the single-reference formalism for each of these geometries, the T1 diagnostic was computed[17] and verified to be ≤ 0.02 in all cases. We computed the energy of each strained conformer relative to the minimized ground-state conformer for each level of theory studied and compared the mean absolute error (MAE) and root-mean-squared error (RMSE) of these predictions for each method.

Benchmarking Methods. We surveyed a wide variety of computational methods: 2 post-Hartree–Fock methods, 17 DFT functionals, 4 composite methods, 4 semiempirical methods, 5 NNPs, and 2 force fields. All calculations were run through ORCA 5.0.3[18] except for: r2SCAN-3c and ωB97X-3c which were run in ORCA 6.0.0; the NNPs, which were run through the Atomic Simulation Environment;[19] and the Sage[20] forcefield, which was run through OpenMM.[21] Unless otherwise specified, DFT calculations were run using the def2-QZVP[22] basis set and wavefunction-based methods were run using the cc-pVQZ[23] basis set. Double hybrid methods

were run with automatically generated auxiliary basis sets using the AutoAux keyword.[24] DLPNO[25] calculations were run with corresponding cc-pVnZ/C auxiliary basis sets and the TightPNO setting applied. CBS extrapolation was performed using the two-point CBS extrapolation method outlined in the ORCA 6 manual (extrapolate_ep1_mdci) using energies from cc-pVTZ and cc-pVQZ calculations.[26,27,28,29]

Timing. First, a few general remarks about timing. Unlike energy-based comparisons, timing comparisons are inherently hardware- and implementation-dependent, and thus are virtually impossible to reproduce with perfect accuracy. As such, the timing benchmarks reported here should be taken as general estimates of the relative speed of different methods, and not as concrete predictions of the amount of time that these calculations will take. Despite this uncertainty, understanding the relative speed of different methods is crucial for designing efficient and scalable computational workflows, and so we here report timing results for all methods under study.

All ORCA calculations were run on 4 CPU cores at the Institute for Cyber-Enabled Research's high-performance-computing cluster at Michigan State University. For DFT and DLPNO-MP2 calculations, 8 GB memory were employed; for DLPNO-CCSD(T) calculations, 96–192 GB of memory were employed. Owing to the complexities of scheduling and compute availability, several different types of nodes were employed: the vast majority of calculations were run with Intel Xeon E5-2680 CPUs (2733 calculations), but a small number of calculations were run with Intel Xeon E7-8867 CPUs (40 calculations), Intel Xeon Gold 6148 CPUs (8 calculations), or AMD EPYC 7763 CPUs (3 calculations). This is anticipated to have a minimal impact on the reported results, since each observed timing datapoint is the average of 153 individual calculations. For all calculations run through ORCA, runtimes correspond to the total elapsed times reported by ORCA.

To test whether running some newer functionals in ORCA (6.0.0) would lead to substantially faster results and prevent fair comparisons of timing, we ran all r2SCAN-3c calculations in both ORCA 6.0.0 and ORCA 5.0.3. We found the calculations run in ORCA 6.0.0 were indeed slightly faster, but that the effect was inconsistent and small relative to the timing differences discussed in this study: calculations run in ORCA 6.0.0 completed in 91±40% of the time that analogous calculations run in ORCA 5.0.3 took to complete. Given that the methods discussed in this study span approximately 17 orders of magnitude in speed, we do not anticipate that the difference between ORCA 5.0.3 and ORCA 6.0.0 will impact our conclusions.

NNP calculations and the Sage forcefield were run on a 2023 Macbook Pro with 11 Apple M3 Pro CPU cores. For NNPs, runtimes were quantified by recording the time to call get_potential_energy() in the Atomic Simulation Environment from an already initialized ase.Calculator object. For Sage, runtime was quantified by recording the time to (1) generate an openmm.State object from an already initialized openmm.Context object and (2) call getPotentialEnergy() from this object. We recognize that this benchmark likely underestimates the speed of Sage in the context of molecular dynamics; however, since Sage is already the fastest method studied here, we did not investigate further speedups. We did not investigate GPU- or TPU-based hardware acceleration in this paper, but further speedups are certainly possible for Sage, all ML-based methods, xTB,[30] and many DFT methods.[31,32]

Results

Geometries and Energies in the Wiggle150 Set. We first investigated the geometries produced by our metadynamics-based workflow. We compared the variation in internal coordinates in our conformer set ("Wiggle150") to that in the large Folmsbee–Hutchison conformer set

("Folmsbee"), in each case comparing the internal coordinate in a given molecule to its corresponding value in the lowest-energy conformer (Figure 1B). While the vast majority of bonds, angles, and dihedrals in the Folmsbee set were almost identical to those in the lowest-energy conformer, the Wiggle150 set showed huge variance: bond lengths varied by up to 0.1 Å from equilibrium, angles varied up to 20º from equilibrium, and many dihedrals covered the entire range of possible values. This structural diversity leads to dramatically more strained structures: while 94% of the Folmsbee set is within 5 kcal/mol of the ground state, the average Wiggle150 conformer is 103 kcal/mol above the ground state. Overall, the metadynamics-based approach used here led to much greater structural diversity than is typically observed in conformer datasets, making this a particularly challenging benchmark.

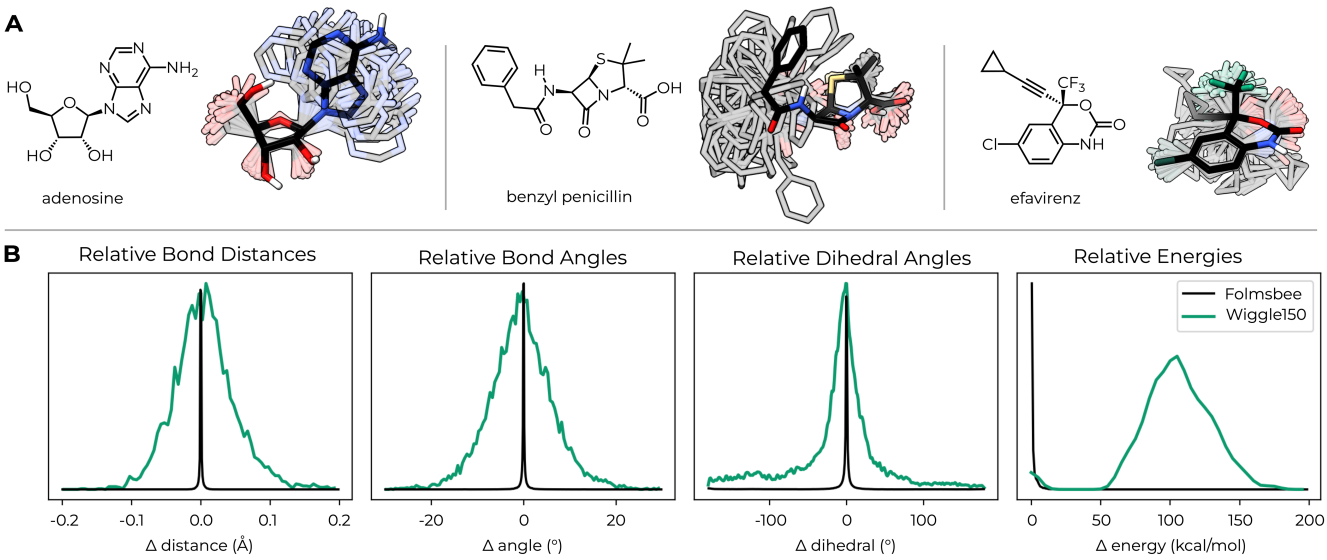

Figure 1. (A) The molecules contained in Wiggle150 and a visualization of the conformers studied. The ground-state conformer is opaque, while the 50 high-energy conformers are translucent. (B) Histograms comparing the bond lengths, angles, dihedral angles, and energies of Wiggle150 as compared to those in the Folmsbee set.

Evaluating Different Methods on Wiggle150. Despite the difficulty of this benchmark set, many high-quality computational methods performed well (Figure 2, Table 1). Numerous methods achieved a MAE under 1 kcal/mol ("chemical accuracy"), including DLPNO-MP2, DSD-PBEP86, and several range-separated-hybrid density functionals.

Excluding double-hybrid functionals, the best-performing functionals all come from Mardirossian and Head-Gordon's work on combinatorial optimization of density functionals. The most accurate functional is the range-separated-hybrid meta-GGA functional ωB97M-V, which uses the Vydrov–van Voorhis non-local dispersion correction.[33,34] Matching results obtained by Martin and co-workers,[35] replacing the VV10 correction with the simpler D3BJ correction leads to a slight decrease in accuracy, although the resulting functional still has the third-lowest MAE of all non-double-hybrid functionals studied. Older global hybrid functionals like wB97X-D3, M06-2X, PBE0, and B3LYP all perform somewhat worse, with PBE0 giving the most consistent performance.

The results of this study support the commonly held idea that ascending the "Jacob's Ladder" of increasing DFT complexity will lead to improved performance: in general, hybrid DFT functionals performed better than non-hybrid/"pure" DFT functionals, and meta-GGA functionals performed better than GGA functionals among both pure and hybrid functionals. Nevertheless, the best-performing pure meta-GGA functionals—r2SCAN, M06-L, and B97M-D4—all performed about as well as common hybrid functionals like M06-2X, B3LYP, and PBE0, demonstrating that pure functionals can give good performance where hybrid DFT is impossible or impractically slow.

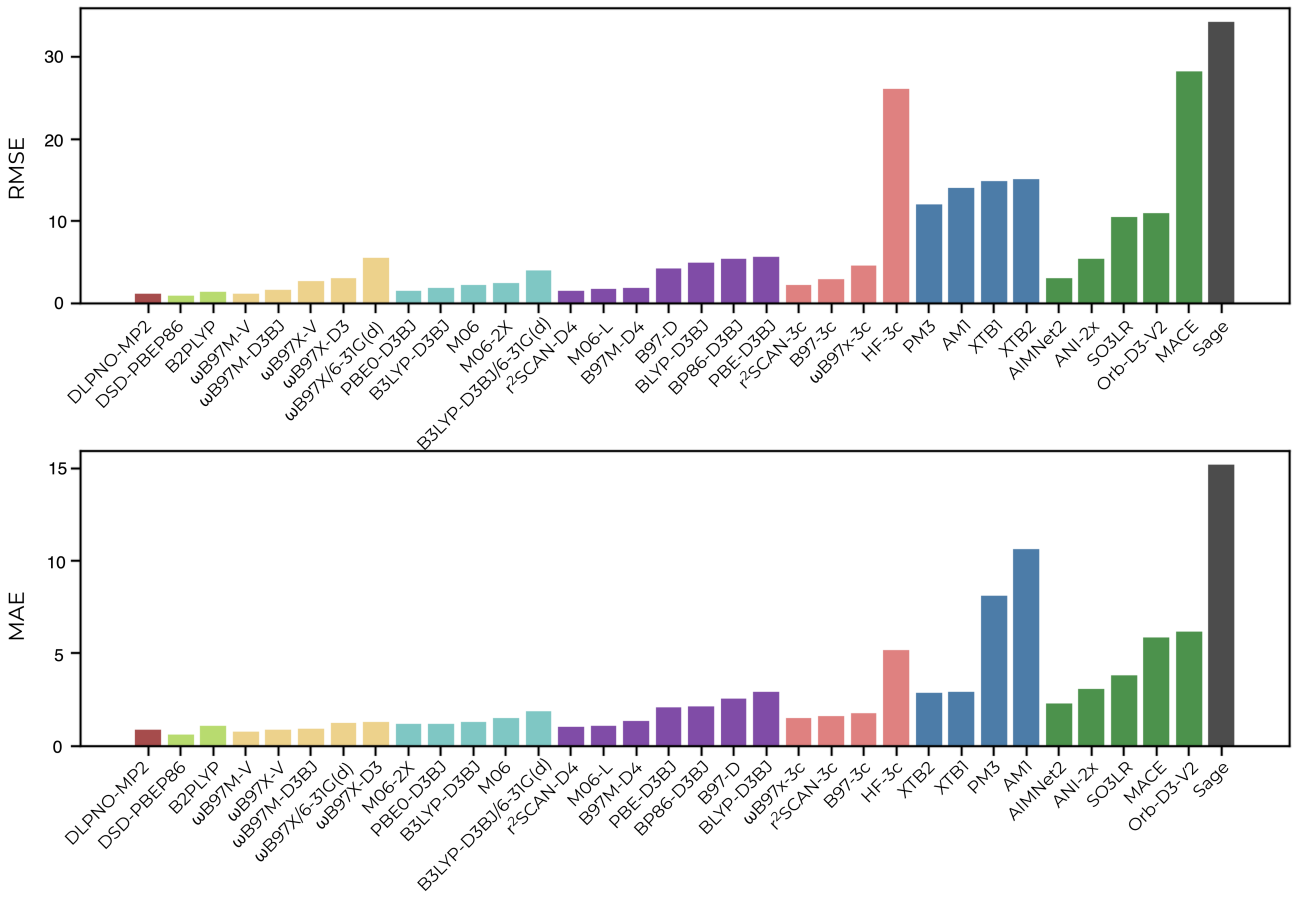

Figure 2. A comparison of the RMSE (top) and MAE (bottom) for all computational methods across all 150 conformers. Methods color coded based on position along "Jacob's Ladder."

Methods using smaller basis sets fared substantially worse on this benchmark set. While full evaluation of basis-set effects on Wiggle150 is outside the scope of this work, we scored two popular double-ζ DFT methods, B3LYP-D3BJ/6-31G(d) and wB97X/6-31G(d), and found that these methods gave markedly worse performance than their quadrupole-ζ congeners. Similarly, the r2SCAN-3c and ωB97X-3c composite methods were markedly worse than their non-composite counterparts. This contrasts with results on GMTKN55, where r2SCAN-3c outperforms r2SCAN-D4/def2-QZVP on relative conformer energies.[36] We attribute this to the high variation in bond length, which likely introduces a small amount of bond breaking into the benchmark; double-ζ

benchmark sets are known to be inaccurate for thermochemistry,[37] so their poor performance here is unsurprising.

In contrast to the overall decent performance of DFT and wavefunction methods, semiempirical methods performed poorly on Wiggle150. We studied 4 semiempirical methods: Grimme and coworkers' semiempirical methods, GFN1-xTB and GFN2-xTB, and the older methods AM1 and PM6. When ranked in terms of MAE, newer methods outcompete older methods: GFN2-xTB > GFN1 > PM6 > AM1. However, when ranked in terms of RMSE, the order is exactly reversed—AM1 performs best, while GFN2-xTB performs worst. These data suggest that the additional complexity of newer semiempirical methods can lead to unreliable performance on strained structures very different from the structures these methods were optimized on.

The 5 NNPs surveyed gave divergent performances. The models trained on materials datasets run with plane-wave PBE calculations, MACE-MP-0[38] and ORB-D3-V2,[39] performed poorly. This is unsurprising: their training data contains few complex organic molecules like the ones shown here and few far-from-equilibrium structures, vividly illustrating how current NNPs struggle to extrapolate beyond their training data with good quantitative accuracy. More surprising is the poor performance of SO3LR,[40] which was trained on 3.5M structures computed at the PBE0+MBD level of theory with numerical atom-centered orbitals in FHI-aims. SO3LR (RMSE: 10.5 kcal/mol) dramatically underperforms the reference PBE0-D3BJ/def2-QZVP results (RMSE: 1.50 kcal/mol), suggesting that either more data or a different architecture is needed to accurately describe these strained molecules.

The remaining two NNPs, ANI-2x[41] and AIMNet2,[42] gave very reasonable results: in particular, AIMNet2 gives a comparable RMSE to many DFT methods (e.g. wB97X-V, B97M-D4, B97-3c),

making it a compelling alternative to conventional quantum-chemical calculations. The accuracy of AIMNet2 is still noticeably lower than the accuracy of the underlying wB97M-D3BJ training data on this benchmark, implying that further gains in the accuracy of this NNP are possible without needing to further increase the level of theory employed for training. The large size of AIMNet2 training dataset (c. 20M) may play a role in the model's observed robustness.

The two forcefields studied here both gave dismal results. The failure of Sage, a conventional forcefield similar to the Amber forcefields, is unsurprising: 2017 study by Kanal and co-workers[43] argued that commonly used forcefields "should not be trusted" for conformer ranking, and that their predictions are "wholly unreliable for conformer screening." In contrast, the poor performance of GFN-FF is more unexpected: in their initial publication describing GFN-FF,[44] Spicher and Grimme report that GFN-FF is "on par with some dispersion-corrected DFT methods" at describing relative conformer energies. Our results show that this is not true for the high-energy conformers studied here.

Table 1: Overview of methods, errors, and average compute time over the Wiggle150 set.

Method	**RMSE** (kcal/mol)	**MAE** (kcal/mol)	**Avg. Time** (s)
DLPNO-CCSD(T)	0.47	0.32	212000
DLPNO-MP2	1.12	0.86	15400
DSD-PBEP86	0.89	0.59	2580
B2PLYP	1.42	1.09	2510
ωB97M-V	1.17	0.77	2008
ωB97X-V	2.69	0.88	1930
ωB97M-D3BJ	1.58	0.94	1970
ωB97X/6-31G(d)	5.49	1.25	180
ωB97X-D3	2.98	1.27	1870
M06-2X	2.40	1.16	1460

PBE0-D3BJ	1.50	1.18	1350
B3LYP-D3BJ	1.83	1.29	1380
M06	2.22	1.52	1450
B3LYP-D3BJ/631G(d)	3.98	1.85	105
r^2SCAN-D4	1.49	1.04	345
M06-L	1.71	1.10	350
B97M-D4	1.82	1.35	353
PBE-D3BJ	5.63	2.06	245
BP86-D3BJ	5.39	2.15	256
BLYP-D3BJ	4.96	2.89	249
B97-D	4.22	2.55	263
ωB97x-3c	4.59	1.50	377
r^2SCAN-3c	2.17	1.61	57.6
B97-3c	2.94	1.77	43.2
HF-3c	26.1	5.15	7.27
GFN2-xTB	15.1	2.87	3.19
GFN1-xTB	14.9	2.90	2.63
PM3	12.1	8.13	1.23
AM1	14.1	10.6	1.28
AIMNet2	3.05	2.31	0.013
ANI-2X	5.40	3.05	0.008
SO3LR	10.5	3.81	2.26
Orb-V2-D3	11.0	6.17	0.0811
MACE-MP-0	28.2	5.83	0.124
Sage 2.2.1	34.2	15.2	0.003
GFN-FF	57.5	41.4	2.63

The Pareto Frontier of Computational Methods. Many of the applications for which good performance on non-equilibrium structures is required, like automated reaction-path exploration and various molecular dynamics–based workflows, also require large numbers of calculations to be run. We compared the speed and accuracy of different methods, with the goal of identifying

points along the Pareto frontier suitable for various use cases. The resulting plots of error vs runtime are shown below (Figure 3).

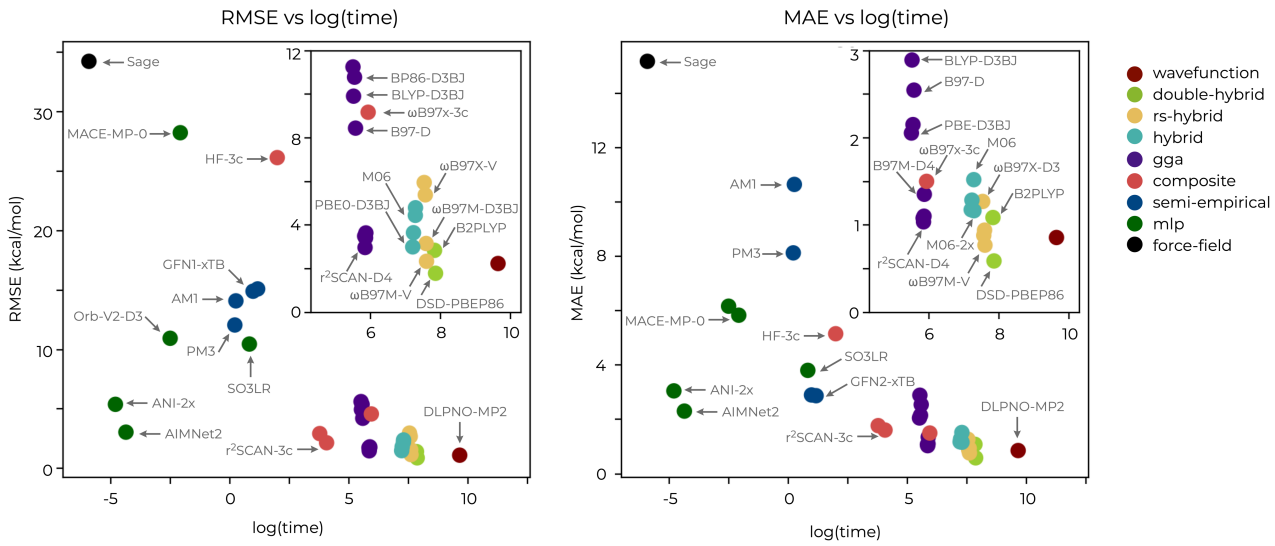

Figure 3. A comparison of the RMSE (left) and MAE (right) versus the log-scaled time in seconds for all computational methods across the Wiggle150 set.

In cases where maximal accuracy is needed, we recommend DSD-PBEP86 or wB97M-V. These methods outperform other hybrid DFT methods with a minimal increase in computational cost: in modern quantum chemistry software, switching to other hybrid functionals like M06-2X or B3LYP offers only minimal savings in time and noticeable decreases in accuracy. If better performance is desired, pure functionals can be used—r2SCAN and M06-L are approximately 5 times faster than top range-separated-hybrid functionals, with slightly increased errors. (The systems investigated in this study are relatively small, and the speedup possible with pure functionals is likely to increase for larger systems).

Among low-cost DFT methods, r2SCAN-3c stands out (as assessed by RMSE). The most commonly used strategy for reducing the cost of DFT simulations is to employ a double-ζ basis

set, often 6-31G(d). Relative to r2SCAN-3c, both B3LYP-D3BJ/6-31G(d) and ωB97X/6-31G(d) give worse accuracy and worse timing, and should be avoided for production use. We note that this study has not comprehensively surveyed many basis-set effects; we leave this substantial task to future work, with the observation that many commonly used basis sets may not be Pareto-efficient.[45,46]

If faster methods than r2SCAN-3c are desired, AIMNet2 is the best choice by far. AIMNet2 offers comparable accuracy to B97-3c while running approximately eight orders of magnitude faster—faster than every semiempirical method, and even faster than GFN-FF. This is dramatically different from the results reported by Folmsbee and Hutchison in their 2021[47] study: in that study, the ANI-2x ML model was similar to GFN1-xTB and GFN2-xTB both in terms of cost and accuracy. In our hands, both ANI-2x and AIMNet2 are significantly faster than xTB-based methods, and AIMNet2 is substantially more accurate, particularly in terms of RMSE. (We note that the present study did not employ hardware acceleration, and that the ML methods used here are likely to be even faster when GPUs are used).

Conclusions

In this study, we generated a benchmark set of strained non-equilibrium conformers of organic small molecules and assessed the ability of various computational methods to predict the relative energies of these conformers. At a high level, the conclusions are straightforward: very good performance is possible on this test set, but high-quality DFT functionals and large basis sets must be employed. Reducing the complexity of the functional or the size of the basis set leads to increased errors, and different balances between speed and accuracy can be achieved with various combinations of method, basis set, and corrections. While some variant of this conclusion is to be

expected from any DFT benchmarking paper, we feel that the small size and considerable difficulty of Wiggle150 makes it a valuable addition to the computational chemistry canon, and we anticipate that this benchmark will prove useful in guiding the creation of future generations of density functionals and NNPs.

Our results also show the impact of recent advances in the development of density functionals. Many of the Pareto-optimal DFT methods are quite new: ωB97M-V[48] was released in 2016, B97M-V[49] & r2SCAN[50] in 2020, and r2SCAN-3c[36] in 2021. As a result, many commonly used software packages do not contain these methods. While a broader discussion of the dynamics of the scientific software ecosystem is outside the scope of this article, our results highlight the reality that many research labs employ suboptimal methods, reducing the accuracy of the results they generate and the speed of their calculations.[51]

Finally, this benchmarking illustrates the remarkable progress made by NNPs over the past several years. The strained molecules studied in Wiggle150 might reasonably have been expected to serve as a "poison" set[52] for machine-learning-based methods, given how few NNPs include structures like this in their training set—instead, NNPs like AIMNet2 and ANI-2x performed very well, approaching in some cases even exceeding the performance of dispersion-corrected DFT methods with quadruple-ζ basis sets. Given that improvement in NNPs continues to proceed at a rapid pace, and that considerable improvement is possible purely from scaling existing architectures to larger datasets,[53] the present authors find it colorable that most quantum mechanical workflows will one day shift to be powered by NNPs.

AUTHOR INFORMATION

Corresponding Author

Joseph J. Gair – Department of Chemistry, Michigan State University, Lansing MI 48824

ORCID 0000-0002-2139-4702 email joegair@msu.edu

Corin C. Wagen – Rowan Scientific, Boston MA 02134 ORCID 0000-0003-3315-3524 email corin@rowansci.com

Conflicts of Interest

C.C.W is a co-founder of the Rowan Scientific Corporation, which produces and commercializes the Rowan platform used to perform some of the calculations described herein.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. ‡These authors contributed equally. (match statement to author names with a symbol)

Funding Sources

This research was supported in part through computational resources provided by the Institute for Cyber-Enabled Research at Michigan State University.

ACKNOWLEDGMENT

C.C.W. acknowledges Arien Wagen and Jonathon Vandezande for helpful discussions and assistance in the preparation of this manuscript. J.J.G acknowledges Mitchell Maday and Paul Reed for technical assistance with MSU HPCC.

REFERENCES

(1) Frances Arnold [@francesarnold]. *You can't always get what you want...but you get what you screen for #directedevolution*. Twitter. https://x.com/francesarnold/status/859998036319117312 (accessed 2025-01-05).

(2) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A Look at the Density Functional Theory Zoo with the Advanced GMTKN55 Database for General Main Group Thermochemistry, Kinetics and Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2017**, *19* (48), 32184–32215. https://doi.org/10.1039/C7CP04913G.

(3) Korth, M.; Grimme, S. "Mindless" DFT Benchmarking. *J. Chem. Theory Comput.* **2009**, *5* (4), 993–1003. https://doi.org/10.1021/ct800511q.

(4) Řezáč, J.; Riley, K. E.; Hobza, P. S66: A Well-Balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.* **2011**, *7* (8), 2427–2438. https://doi.org/10.1021/ct2002946.

(5) Brauer, B.; Kesharwani, M. K.; Kozuch, S.; Martin, J. M. L. The S66x8 Benchmark for Noncovalent Interactions Revisited: Explicitly Correlated Ab Initio Methods and Density Functional Theory. *Phys. Chem. Chem. Phys.* **2016**, *18* (31), 20905–20925. https://doi.org/10.1039/C6CP00688D.

(6) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering Chemistry with an Ab Initio Nanoreactor. *Nature Chem* **2014**, *6* (12), 1044–1048. https://doi.org/10.1038/nchem.2099.

(7) Chang, A. M.; Meisner, J.; Xu, R.; Martínez, T. J. Efficient Acceleration of Reaction Discovery in the Ab Initio Nanoreactor: Phenyl Radical Oxidation Chemistry. *J Phys Chem A* **2023**, *127* (45), 9580–9589. https://doi.org/10.1021/acs.jpca.3c05484.

(8) Barroso-Luque, L.; Shuaibi, M.; Fu, X.; Wood, B. M.; Dzamba, M.; Gao, M.; Rizvi, A.; Zitnick, C. L.; Ulissi, Z. W. Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models. arXiv October 16, 2024. https://doi.org/10.48550/arXiv.2410.12771.

(9) Herr, J. E.; Yao, K.; McIntyre, R.; Toth, D. W.; Parkhill, J. Metadynamics for Training Neural Network Model Chemistries: A Competitive Assessment. *The Journal of Chemical Physics* **2018**, *148* (24), 241710. https://doi.org/10.1063/1.5020067.

(10) Deng, B.; Choi, Y.; Zhong, P.; Riebesell, J.; Anand, S.; Li, Z.; Jun, K.; Persson, K. A.; Ceder, G. Overcoming Systematic Softening in Universal Machine Learning Interatomic Potentials by Fine-Tuning. arXiv May 11, 2024. https://doi.org/10.48550/arXiv.2405.07105.

(11) Fu, X.; Wu, Z.; Wang, W.; Xie, T.; Keten, S.; Gomez-Bombarelli, R.; Jaakkola, T. Forces Are Not Enough: Benchmark and Critical Evaluation for Machine Learning Force Fields with Molecular Simulations. *Transactions on Machine Learning Research* **2023**.

(12) *Rowan Documentation*. Rowan Documentation. https://docs.rowansci.com (accessed 2025-01-05).

(13) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2019**, *15* (5), 2847–2862. https://doi.org/10.1021/acs.jctc.9b00143.

(14) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15* (3), 1652–1671. https://doi.org/10.1021/acs.jctc.8b01176.

(15) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J Chem Inf Model* **2015**, *55* (12), 2562–2574. https://doi.org/10.1021/acs.jcim.5b00654.

(16) Wang, S.; Witek, J.; Landrum, G. A.; Riniker, S. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *J. Chem. Inf. Model.* **2020**, *60* (4), 2044–2058. https://doi.org/10.1021/acs.jcim.0c00025.

(17) Lee, T. J.; Taylor, P. R. A Diagnostic for Determining the Quality of Single-Reference Electron Correlation Methods. *Int. J. Quantum Chem.* **2009**, *36* (S23), 199–207. https://doi.org/10.1002/qua.560360824.

(18) Neese, F. Software Update: The ORCA Program System—Version 5.0. *WIREs Comput Mol Sci* **2022**, *12* (5), e1606. https://doi.org/10.1002/wcms.1606.

(19) Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.; Jennings, P. C.; Bjerre Jensen, P.; Kermode, J.; Kitchin, J. R.; Leonhard Kolsbjerg, E.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Bergmann Maronsson, J.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schiøtz, J.; Schütt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z.; Jacobsen, K. W. The Atomic Simulation Environment—a Python Library for Working with Atoms. *J. Phys.: Condens. Matter* **2017**, *29* (27), 273002. https://doi.org/10.1088/1361-648X/aa680e.

(20) Boothroyd, S.; Behara, P. K.; Madin, O. C.; Hahn, D. F.; Jang, H.; Gapsys, V.; Wagner, J. R.; Horton, J. T.; Dotson, D. L.; Thompson, M. W.; Maat, J.; Gokey, T.; Wang, L.-P.; Cole, D. J.; Gilson, M. K.; Chodera, J. D.; Bayly, C. I.; Shirts, M. R.; Mobley, D. L. Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *J. Chem. Theory Comput.* **2023**, *19* (11), 3251–3275. https://doi.org/10.1021/acs.jctc.3c00039.

(21) Eastman, P.; Galvelis, R.; Peláez, R. P.; Abreu, C. R. A.; Farr, S. E.; Gallicchio, E.; Gorenko, A.; Henry, M. M.; Hu, F.; Huang, J.; Krämer, A.; Michel, J.; Mitchell, J. A.; Pande, V. S.; Rodrigues, J. P.; Rodriguez-Guerra, J.; Simmonett, A. C.; Singh, S.; Swails, J.; Turner, P.; Wang, Y.; Zhang, I.; Chodera, J. D.; De Fabritiis, G.; Markland, T. E. OpenMM 8: Molecular

Dynamics Simulation with Machine Learning Potentials. *J. Phys. Chem. B* **2024**, *128* (1), 109–116. https://doi.org/10.1021/acs.jpcb.3c06662.

(22) Weigend, F.; Furche, F.; Ahlrichs, R. Gaussian Basis Sets of Quadruple Zeta Valence Quality for Atoms H–Kr. *The Journal of Chemical Physics* **2003**, *119* (24), 12753–12762. https://doi.org/10.1063/1.1627293.

(23) Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *The Journal of Chemical Physics* **1989**, *90* (2), 1007–1023. https://doi.org/10.1063/1.456153.

(24) Stoychev, G. L.; Auer, A. A.; Neese, F. Automatic Generation of Auxiliary Basis Sets. *J Chem Theory Comput* **2017**, *13* (2), 554–562. https://doi.org/10.1021/acs.jctc.6b01041.

(25) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural Triple Excitations in Local Coupled Cluster Calculations with Pair Natural Orbitals. *The Journal of Chemical Physics* **2013**, *139* (13), 134101. https://doi.org/10.1063/1.4821834.

(26) Liakos, D. G.; Neese, F. Improved Correlation Energy Extrapolation Schemes Based on Local Pair Natural Orbital Methods. *J. Phys. Chem. A* **2012**, *116* (19), 4801–4816. https://doi.org/10.1021/jp302096v.

(27) Neese, F.; Valeev, E. F. Revisiting the Atomic Natural Orbital Approach for Basis Sets: Robust Systematic Basis Sets for Explicitly Correlated and Conventional Correlated Ab Initio Methods? *J Chem Theory Comput* **2011**, *7* (1), 33–43. https://doi.org/10.1021/ct100396y.

(28) Zhong, S.; Barnes, E. C.; Petersson, G. A. Uniformly Convergent N-Tuple-ζ Augmented Polarized (nZaP) Basis Sets for Complete Basis Set Extrapolations. I. Self-Consistent Field

Energies. *The Journal of Chemical Physics* **2008**, *129* (18), 184116. https://doi.org/10.1063/1.3009651.

(29) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-Set Convergence of Correlated Calculations on Water. *The Journal of Chemical Physics* **1997**, *106* (23), 9639–9646. https://doi.org/10.1063/1.473863.

(30) Friede, M.; Hölzer, C.; Ehlert, S.; Grimme, S. *Dxtb* —An Efficient and Fully Differentiable Framework for Extended Tight-Binding. *The Journal of Chemical Physics* **2024**, *161* (6), 062501. https://doi.org/10.1063/5.0216715.

(31) Ju, F.; Wei, X.; Huang, L.; Jenkins, A. J.; Xia, L.; Zhang, J.; Zhu, J.; Yang, H.; Shao, B.; Dai, P.; Williams-Young, D. B.; Mayya, A.; Hooshmand, Z.; Efimovskaya, A.; Baker, N. A.; Troyer, M.; Liu, H. Acceleration without Disruption: DFT Software as a Service. *J. Chem. Theory Comput.* **2024**, *20* (24), 10838–10851. https://doi.org/10.1021/acs.jctc.4c00940.

(32) Wang, Y.; Hait, D.; Johnson, K. G.; Fajen, O. J.; Zhang, J. H.; Guerrero, R. D.; Martínez, T. J. Extending GPU-Accelerated Gaussian Integrals in the TeraChem Software Package to f Type Orbitals: Implementation and Applications. *J Chem Phys* **2024**, *161* (17), 174118. https://doi.org/10.1063/5.0233523.

(33) Vydrov, O. A.; Van Voorhis, T. Implementation and Assessment of a Simple Nonlocal van Der Waals Density Functional. *The Journal of Chemical Physics* **2010**, *132* (16), 164113. https://doi.org/10.1063/1.3398840.

(34) Vydrov, O. A.; Van Voorhis, T. Nonlocal van Der Waals Density Functional: The Simpler the Better. *The Journal of Chemical Physics* **2010**, *133* (24), 244103. https://doi.org/10.1063/1.3521275.

(35) Santra, G.; Martin, J. M. L. Some Observations on the Performance of the Most Recent Exchange-Correlation Functionals for the Large and Chemically Diverse GMTKN55 Benchmark. *AIP Conference Proceedings* **2019**, *2186* (1), 030004. https://doi.org/10.1063/1.5137915.

(36) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. r2SCAN-3c: A "Swiss Army Knife" Composite Electronic-Structure Method. *The Journal of Chemical Physics* **2021**, *154* (6), 064103. https://doi.org/10.1063/5.0040021.

(37) Bursch, M.; Mewes, J.; Hansen, A.; Grimme, S. Best-Practice DFT Protocols for Basic Molecular Computational Chemistry**. *Angewandte Chemie* **2022**, *134* (42), e202205735. https://doi.org/10.1002/ange.202205735.

(38) Batatia, I.; Benner, P.; Chiang, Y.; Elena, A. M.; Kovács, D. P.; Riebesell, J.; Advincula, X. R.; Asta, M.; Avaylon, M.; Baldwin, W. J.; Berger, F.; Bernstein, N.; Bhowmik, A.; Blau, S. M.; Cărare, V.; Darby, J. P.; De, S.; Pia, F. D.; Deringer, V. L.; Elijošius, R.; El-Machachi, Z.; Falcioni, F.; Fako, E.; Ferrari, A. C.; Genreith-Schriever, A.; George, J.; Goodall, R. E. A.; Grey, C. P.; Grigorev, P.; Han, S.; Handley, W.; Heenen, H. H.; Hermansson, K.; Holm, C.; Jaafar, J.; Hofmann, S.; Jakob, K. S.; Jung, H.; Kapil, V.; Kaplan, A. D.; Karimitari, N.; Kermode, J. R.; Kroupa, N.; Kullgren, J.; Kuner, M. C.; Kuryla, D.; Liepuoniute, G.; Margraf, J. T.; Magdău, I.-B.; Michaelides, A.; Moore, J. H.; Naik, A. A.; Niblett, S. P.; Norwood, S. W.; O'Neill, N.; Ortner, C.; Persson, K. A.; Reuter, K.; Rosen, A. S.; Schaaf, L. L.; Schran, C.; Shi, B. X.; Sivonxay, E.; Stenczel, T. K.; Svahn, V.; Sutton, C.; Swinburne, T. D.; Tilly, J.; Oord, C. van der; Varga-

Umbrich, E.; Vegge, T.; Vondrák, M.; Wang, Y.; Witt, W. C.; Zills, F.; Csányi, G. A Foundation Model for Atomistic Materials Chemistry. arXiv March 1, 2024. https://doi.org/10.48550/arXiv.2401.00096.

(39) Neumann, M.; Gin, J.; Rhodes, B.; Bennett, S.; Li, Z.; Choubisa, H.; Hussey, A.; Godwin, J. Orb: A Fast, Scalable Neural Network Potential. arXiv October 29, 2024. https://doi.org/10.48550/arXiv.2410.22570.

(40) Kabylda, A.; Frank, J. T.; Dou, S. S.; Khabibrakhmanov, A.; Sandonas, L. M.; Unke, O. T.; Chmiela, S.; Muller, K.-R.; Tkatchenko, A. Molecular Simulations with a Pretrained Neural Network and Universal Pairwise Force Fields. Chemistry October 8, 2024. https://doi.org/10.26434/chemrxiv-2024-bdfr0.

(41) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16* (7), 4192–4202. https://doi.org/10.1021/acs.jctc.0c00121.

(42) Anstine, D.; Zubatyuk, R.; Isayev, O. AIMNet2: A Neural Network Potential to Meet Your Neutral, Charged, Organic, and Elemental-Organic Needs. Chemistry December 20, 2024. https://doi.org/10.26434/chemrxiv-2023-296ch-v3.

(43) Kanal, I. Y.; Keith, J. A.; Hutchison, G. R. A Sobering Assessment of Small-molecule Force Field Methods for Low Energy Conformer Predictions. *Int J of Quantum Chemistry* **2018**, *118* (5), e25512. https://doi.org/10.1002/qua.25512.

(44) Spicher, S.; Grimme, S. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angew Chem Int Ed* **2020**, *59* (36), 15665–15673. https://doi.org/10.1002/anie.202004239.

(45) Gray, M.; Bowling, P. E.; Herbert, J. M. Comment on "Benchmarking Basis Sets for Density Functional Theory Thermochemistry Calculations: Why Unpolarized Basis Sets and the Polarized 6-311G Family Should Be Avoided." *J. Phys. Chem. A* **2024**, *128* (36), 7739–7745. https://doi.org/10.1021/acs.jpca.4c00283.

(46) Behara, P. K.; Jang, H.; Horton, J. T.; Gokey, T.; Dotson, D. L.; Boothroyd, S.; Bayly, C. I.; Cole, D. J.; Wang, L.-P.; Mobley, D. L. Benchmarking Quantum Mechanical Levels of Theory for Valence Parametrization in Force Fields. *J. Phys. Chem. B* **2024**, *128* (32), 7888–7902. https://doi.org/10.1021/acs.jpcb.4c03167.

(47) Folmsbee, D.; Hutchison, G. Assessing Conformer Energies Using Electronic Structure and Machine Learning Methods. *Int J of Quantum Chemistry* **2021**, *121* (1), e26381. https://doi.org/10.1002/qua.26381.

(48) Mardirossian, N.; Head-Gordon, M. ω B97M-V: A Combinatorially Optimized, Range-Separated Hybrid, Meta-GGA Density Functional with VV10 Nonlocal Correlation. *The Journal of Chemical Physics* **2016**, *144* (21), 214110. https://doi.org/10.1063/1.4952647.

(49) Najibi, A.; Goerigk, L. DFT -D4 Counterparts of Leading META- Generalized-gradient Approximation and Hybrid Density Functionals for Energetics and Geometries. *J Comput Chem* **2020**, *41* (30), 2562–2572. https://doi.org/10.1002/jcc.26411.

(50) Furness, J. W.; Kaplan, A. D.; Ning, J.; Perdew, J. P.; Sun, J. Accurate and Numerically Efficient r^2 SCAN Meta-Generalized Gradient Approximation. *J. Phys. Chem. Lett.* **2020**, *11* (19), 8208–8215. https://doi.org/10.1021/acs.jpclett.0c02405.

(51) Armstrong, G. *Five years of polling the computational chemistry community*. https://blogs.nature.com/thescepticalchymist/2014/11/five-years-of-polling-the-computational-chemistry-community.html.

(52) Gould, T.; Dale, S. G. Poisoning Density Functional Theory with Benchmark Sets of Difficult Systems. *Phys. Chem. Chem. Phys.* **2022**, *24* (11), 6398–6403. https://doi.org/10.1039/D2CP00268J.

(53) Frey, N. C.; Soklaski, R.; Axelrod, S.; Samsi, S.; Gómez-Bombarelli, R.; Coley, C. W.; Gadepally, V. Neural Scaling of Deep Chemical Models. *Nat Mach Intell* **2023**, *5* (11), 1297–1305. https://doi.org/10.1038/s42256-023-00740-3.

Table of contents graphic